Successful explanations start with accurate descriptions

Accepted to European Journal of Personality (14th February 2018)

Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions

Anne Seeboth *

Department of Psychology, University of Edinburgh

René Mõttus *

Department of Psychology, Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh

Institute of Psychology, University of Tartu

* Department of Psychology, University of Edinburgh
7 George Square
EH8 9JZ Edinburgh, UK
s1311783@sms.ed.ac.uk or rene.mottus@ed.ac.uk

Author's Note

The analyses in this work are based on data from the National Child Development Study (NCDS). The authors are grateful to the Centre for Longitudinal Studies (CLS), UCL Institute of Education for the use of these data and to the UK Data Service for making them available. However, neither CLS nor the UK Data Service bear any responsibility for the analysis or interpretation of these data.

Abstract

Personality-outcome associations, typically represented using the Big Five personality domains, are ubiquitous, but often weak and possibly driven by the constituents of these domains. We hypothesized that representing the associations using personality questionnaire items (as markers for personality nuances) could increase prediction strength. Using the National Child Development Study (N = 8,719), we predicted 40 diverse outcomes from both the Big Five domains and their 50 items. Models were trained (using penalized regression) and applied for prediction in independent sample partitions (with 100 permutations). Item-models tended to out-predict Big Five-models (explaining on average 30% more variance), regardless of outcomes' independently-rated breadth *versus* behavioral specificity. Moreover, the predictive power of Big Five domains *per se* was at least partly inflated by the unique variance of their constituent items, especially for generally more predictable outcomes. Removing the Big Five variance from items marginally reduced their predictive power. These findings are consistent with the possibility that the associations of personality with outcomes often pertain to (potentially large numbers of) specific behavioral, cognitive, affective and motivational characteristics represented by single questionnaire items rather than to the broader (underlying) traits that these items are ostensibly indicators of. This may also have implications for personality-based interventions.

Keywords: machine learning; personality; items; outcomes; validity

Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions

Among the central questions of personality research are associations of personality characteristics with life outcomes, defined as phenomena that could potentially be influenced by personality (Mõttus, 2016). From an applied perspective, understanding these associations may allow for the identification of people at risk of negative life outcomes, such as unemployment or diabetes, and for the discovery of potentially modifiable risk factors related to these outcomes (e.g., specific academic difficulties or health-related life-style aspects). From a psychological-theoretical point of view, delineating these associations allows for understanding where, to what extent and perhaps even how personality may play out in people's lives. We argue that achieving these aims can benefit from the most accurate possible description of how personality is associated with outcomes, even if this means very nuanced patterns of associations—in which case this very realization is telling.

Most commonly, personality-outcome associations are represented using the five broad domains of the Five Factor Model (FFM; McCrae & John, 1992) or the Big Five (Goldberg, 1990): Conscientiousness, Extraversion, Agreeableness, Neuroticism and Openness. Based on this representation, the associations are ubiquitous. However, we argue, they tend to be relatively weak and are often unspecific, such that even very different outcomes are related to similar trait combinations. One potential way to improve the accuracy and specificity of the associations may be to investigate them using a larger set of narrower, more specific personality traits than the Big Five. Although personality facets (e.g., McCrae & Costa, 2010; Soto & John, 2017) can be useful for this, recent work suggests that even single personality questionnaire items contain

unique information that may be lost in aggregation (e.g., Mõttus, Kandler, Bleidorn, Riemann & McCrae, 2017; Mõttus, Sinick, Terracciano et al., under review). In this study, therefore, we explore the usefulness of single questionnaire items as *personality markers* (analogously to genetic markers in molecular genetics research) in accounting for the variance in 40 outcomes representing individual differences in a variety of life domains. Specifically, we focus on items' predictive accuracy for these outcomes in comparison to the Big Five traits and address the possibility that Big Five-outcome associations tend to be at least partly driven by the items that happen to be included in them rather than (only) by the latent traits purportedly underlying the Big Five scores.

Representing Personality-Outcome Associations Using Domains, Facets and Items

Domains

The Big Five personality domains are robustly associated with a wide range of broad life outcomes, such as physical and mental health (e.g., Goodwin & Friedman, 2006) or success in education, career and relationships (e.g., Damian, Su, Shanahan, Trautwein & Roberts, 2015; Poropat, 2009; Roberts, Kuncel, Shiner, Caspi & Goldberg, 2007). In addition, several of the domains are linked to more specific outcomes, such as drinking or smoking (Malouff, Thorsteinsson, Rooke & Schutte, 2007; Malouff, Thorsteinsson & Schutte, 2006), relationship satisfaction (Malouff, Thorsteinsson, Schutte, Bhullar & Rooke, 2010), volunteering (Carlo, Okun, Knight & de Guzman, 2005), exercising (Rhodes & Smith, 2006) or voting choices (Vecchione et al., 2011). Although ubiquitous, the associations are often modest in strength, with most effect sizes (*r*) well below .30 and often even below .20 or .10 (especially in larger samples). And because the Big Five traits are inter-correlated, the unique associations are often smaller still (e.g., Laidra, Pullmann, & Allik, 2007). Of course, individual effect sizes are expected to be modest in psychology—and they are generally small (Richard, Bond & Stokes-Zoota, 2003)—because any behavioral phenomenon is likely to be linked with a huge number of causal factors. There is nothing inherently wrong with small effects. However, we argue that personality-outcome associations may often be at least somewhat stronger than can be estimated by means of the Big Five domains, and that quantifying the full magnitude of these associations when and where they are, in fact, stronger than is observed based on the Big Five can be useful for advancing our understanding of how personality intersects with life outcomes.

Also, while the Big Five domains do relate to a wide range of broad and narrow life outcomes, the associations are often rather unspecific. Positive outcomes such as higher educational level and income, relationship quality, lack of antisocial behavior, healthy life-style or longevity tend to be associated with low Neuroticism on one side and high Conscientiousness and Agreeableness on the other, and they often share smaller links with high Extraversion and/or Openness (Damian et al, 2015; Graham et al., 2017; Jones, Miller & Lynam, 2011; Malouff et al, 2010; Mõttus et al., 2012). The opposite pattern tends to characterize negative outcomes. Because this pattern corresponds to the social desirability of the Big Five domains (Allik et al., 2010), it suggests that the domains may be to a substantial extent associated with the general valence in outcomes rather than with their specific aspects—the aspects that make even similarly-valenced outcomes distinct. To the extent that this applies, it limits the informativeness of the associations beyond suggesting that, perhaps accurately, positive things hang together.

Facets

Albeit far less frequently, personality-outcome associations have also been investigated at narrower levels of the personality hierarchy. Each Big Five domain, for instance, has been suggested to be made up of six facets (e.g., McCrae & Costa, 2010), although this particular facet-model is not necessarily based on empirical work (McCrae & Costa, 2008) and there exist alternative facet-models for the Big Five domains (e.g., Soto & John, 2017). Besides contributing to the domain particular facets are intended to measure, they also capture unique variance that is not shared with other facets of the same domain and that corresponds to distinct etiological mechanisms (e.g., Jang, McCrae, Angleitner, Riemann & Livesley, 1998). This facet-specific variance is also associated with life outcomes, and therefore including facets into prediction models of these outcomes can increase their predictive power (e.g., Anglim & Grant, 2016; Christiansen & Robie, 2011; Paunonen & Ashton, 2001).

In addition, facet-outcome links are likely to be more specific than those based on the Big Five domains. Therefore, using facets can entail not only more accurate but also more diverse representations of personality-outcome associations. For example, both Body Mass Index (BMI) and aggressiveness have small positive associations with Neuroticism, but the two outcomes differ substantially in their links with Neuroticism facets; for the former, the association only pertains to the Impulsivity facet, whereas the latter has the strongest link with the Angry Hostility facet (Jones et al., 2011; Sutin, Ferrucci, Zonderman & Terracciano, 2011). When this applies, arguably, personality-outcome associations should be interpreted at the level of facets and not be generalized to domains at all (Mõttus, 2016).

Personality-Outcome Associations from Nuances (Items)

Facets may not be the most specific personality characteristics. Recently, McCrae (2015) suggested that the hierarchy of personality traits extends even below facets, to narrow personality characteristics that he called *nuances*. Due to a lack of proper classification, nuances have thus far been operationalized as individual personality questionnaire items. Nuances contain unique variance that is not shared with Big Five domains and their facets and tends to have trait-like properties of cross-rater agreement (Mõttus, McCrae, Allik & Realo, 2014; Mõttus, Kandler et al., 2017; Mõttus et al., under review), stability over time and a non-zero level of heritability (Mõttus, Kandler et al., 2017; Mõttus et al., under review). Also, nuances often display varying gender-differences and age-trends from the domains and facets that they are intended to be indicators of (Mõttus et al., 2015; Mõttus et al., under review). Some of this unique variance is filtered out when items are aggregated into domain and facet scores but it may be useful for outcome prediction.

For example, BMI is associated with the unique variance of items related to overeating (Vainik, Mõttus, Allik, Esko & Realo, 2015) and giving up on self-improvement programs (Mõttus, Kandler et al., 2017), among a range of other items, and such associations tend to replicate across samples from different countries (Mõttus et al., under review). Likewise, items' residual variance (after adjusting for the variance of the Big Five traits and their facets) has meaningful associations with people's interests in various life domains (Mõttus, Kandler et al., 2017). For other outcomes, nuance-specific variance may appear less relevant: Mõttus, Kandler and colleagues (2017) found that item residuals were not significantly correlated with life satisfaction after Bonferroni correction for the number of correlations they had investigated.

In fact, there may be systematic regularities in where items' unique variance provides incremental predictive value. It has been suggested that for optimal prediction, predictors and outcomes should be matched in terms of their breadth (Asendorpf et al., 2016; Wittmann, 1988). If so, items should be better predictors of more specific outcomes (if outcome-relevant items have been included in the personality measure) whereas composite traits that aggregate multiple behavioral, affective and cognitive tendencies should out-predict items for broader outcomes that also aggregate the cumulative results of multiple behaviors, thoughts and feelings (Mõttus, Kandler et al., 2017).

On the other hand, a finding that items' unique variance does not *significantly* correlate with an outcome in a given sample (Mõttus, Kandler et al., 2017) does not necessarily mean that it does not relate to the outcome at all. For instance, residuals of some items may well be associated with life satisfaction, but their *individual* effect sizes may be too small to reach significance, especially when stringent significance criteria (adjusted for large numbers of associations being tested) and not very large samples are used. *Cumulatively*, however, outcome prediction models including a number of questionnaire items may predict outcomes better than the Big Five domains even when individual links between the items' unique variance and outcomes are weak. We appreciate that this reasoning may seem surprising to some readers, especially given the currently widely prevalent concerns around replicability—one has to be mindful of the dangers of over-fitting models to data, either accidentally or deliberately (Yarkoni & Westfall, 2017). But rest assured: we take this danger very seriously. In order to illustrate our thinking, it may help to draw a parallel with molecular genetics.

A Parallel with Molecular Genetics

Geneticists have come to realize that complex phenotypes tend to be linked to hundreds or thousands of genetic variants each conferring only a tiny effect, rather than to a few genetic variants with strong effects that can be easily detected; this is known as the Fourth Law of Behavior Genetics (Chabris, Lee, Cesarini, Benjamin & Laibson, 2015). Such phenotypes—that is, most if not all phenomena personality psychologists are striving to learn about—are called *polygenic*. To study the individually small but potentially numerous genetic effects, researchers atheoretically link millions of genetic markers, single nucleotide polymorphisms (SNPs) designating allelic variations in specific regions of the genome, to a given phenotype, a method known as genome-wide association studies (GWAS; Hirschhorn & Daly, 2005). Given large enough samples (which means tens of thousands of participants or even more), predictive models built on the basis of the GWAS results can often explain substantial amounts of variance in complex phenotypes such as intelligence (Davies et al., 2015), schizophrenia (Lee et al., 2012), BMI (Locke et al., 2015) or educational attainment (Okbay et al., 2016), even when applied to independent samples of people. Such models can be used to create *polygenic scores*: individuals' estiamted genetic propensities for a given phenotype, derived by summing weighted allelic counts across large numbers of SNPs, with the weights taken from a GWAS carried out in independent samples (for a detailed yet accessible explanation, see Plomin & von Stumm, 2018). In addition to simply allowing for phenotypic prediction from genomic information, GWAS have also started to unravel the multi-faceted biological etiology of complex phenotypes such as BMI (Locke et al., 2015), depression (Major Depressive Disorder Working Group of the PGC, Wray, & Sullivan, 2017) or intelligence (Hill, Davies, McIntosh, Gale, & Deary, 2017).

Importantly, including even very small effects of the genetic markers that, individually, are not statistically significantly associated with the phenotypes—and an overwhelming majority of them are not—typically contributes to the amount of variance explained by polygenic scores (Dudbridge, 2013). For example, SNP-intelligence associations (from a GWAS meta-analysis based on up to about 280,000 people in total) with *p*-values of up to .26 contributed to the prediction of observed intelligence in four independent samples (Savage et al., 2017). This means that even tiny and, by conventional criteria, non-significant effects are often in fact real effects in that they contribute to the predictive signal. Realizing this is important not only because it allows for more accurate predictive models, but also because this tells geneticists something very important about the very nature of the genetic etiology of complex phenotypes: genetically, they are so multiply determined that singling out any one—or even dozens or hundreds—genetic variant(s) as *the gene(s)* for any one phenotype may often not make much sense. For an accessible account of recent progress in GWAS and polygenic scoring research and its major implications for understanding behavioral phenomena (with intelligence as the focal trait) readers are referred to Plomin and von Stumm (2018).

Analogously, just as complex phenotypes are highly polygenic, we suggest that complex outcomes may turn out to be *polynuanced*. In other words, their associations with personality variations may be driven by a large number of specific personality characteristics in addition to, or perhaps sometimes even instead of, a small number of broad "underlying" constructs that composite personality traits ought to represent. To address this possibility, questionnaire items could be used as *personality markers* of (yet unknown) nuances. What is more, full sets of items could be atheoretically linked to outcomes in *questionnaire-wide association studies* (QWAS; the term can be traced to Weiss, Gale, Batty & Deary, 2013, although their rationale for and procedures of QWAS differed from ours). If and when outcomes turn out to be polynuanced, this will have implications for both prediction *per se* and, more generally, for our understanding of how personality intersects with phenomena outside the personality domain. As for prediction, this could suggest that optimal models should be built based on large numbers of narrow personality characteristics (nuances). Such predictions could be called *polynuance scores*, analogously to polygenic scores. Polynuance scores are also similar to empirically-constructed personality scales such as those of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1940) and the California Personality Inventory (CPI; Gough, 1975), with the difference that polynuance scores comprise weighted (by regression coefficients) contributions of items, whereas this is not necessarily the case for scores of empirically-constructed personality scales. Such polynuance scores could be used wherever personality characteristics are used for prediction—for example, for selecting best candidates based on predicted future performance or identification of people at future risk for certain negative outcomes—but they would provide higher predictive value than predictions based on domains and perhaps also facets. As for understanding, in addition to singling out specific aspects of personality that are linked with particular outcomes, investigating the sheer numbers of such links may be informative regarding how personality and outcomes tend to intersect (e.g., *how* polynuanced they tend to be). Furthermore, if outcomes are indeed correlated with ranges of nuances, this may also increase the likelihood that different outcomes correspond to a more varied set of personality profiles than the all-positive-goes-together pattern that often emerges when the Big Five domains are used. If so, there would be comparatively more to personality in

relation to outcomes than a desirable personality tagging along with other desirable characteristics.

Let us be clear: we are not postulating that all outcomes are necessarily polynuanced—they may well not be at all, or only some of them may be. But we do suggest that this is an entirely realistic possibility with potentially broad implications, and that it is therefore something worthwhile exploring. To exactly this end, we systematically compared the degrees to which item-based models could predict a range of outcomes compared to the degrees to which the outcomes could be predicted from the Big Five domains. We also explored the possibility that the Big Five-outcome associations may be driven by the items that had been included in the domain scores.

Methodological Considerations

Adding more predictors to a model tends to increase the amount of variance the model can explain in the sample where the model is fitted, but such a model might perform poorly when applied to a different sample of the same population: this is known as model over-fitting (e.g., Chapman, Weiss & Duberstein, 2016; Yarkoni & Westfall, 2017). Therefore, item-based models may tend to out-predict domain-based models for purely statistical reasons. In order to mitigate this danger, geneticists (and all adopters of machine learning principles) often create (train) the prediction models in one sample and apply (cross-validate) them in independent samples. So could do personality researchers. In fact, the training and cross-validation sample do not necessarily have to be independent in the sense that they are collected by different researchers and/or at different times/sites. A single, sufficiently large sample can be split into two independent partitions, one for model training and the other for validation, and this procedure could be repeated multiple times, which yields a distribution for the parameter of interest such as the amount of variance in an outcome the model can account for in independent groups of people.

In order to increase the likelihood of a model performing well in cross-validation, it is advisable to ensure that its parameters (e.g., regression coefficients) are estimated as well as possible in the training sample. When large numbers of inter-correlated predictors are included in a model, traditional least-square or maximum likelihood regressions may struggle to produce coefficients that yield optimal predictions in independent samples because they are tailored to the idiosyncrasies of the particular sample (i.e., over-fit) and because of possible multi-collinearity among predictors. Increasing the sample size helps with the former, but not the latter. Therefore, it may be useful to employ regularized regression approaches such as ridge (Hoerl & Kennard, 1970), Least Absolute Shrinkage and Selection Operator (LASSO) or elastic net (Tibshirani, 2011; Zou & Hastie, 2005) for training models. These regression methods are designed to deal with large numbers of inter-correlated predictors and yield more parsimonious (compared to more traditional, non-regularized approaches) models that are less prone to over-fitting to start with. Specifically, these approaches penalize regression coefficients by shrinking them towards zero, because this counteracts the natural tendency of regression models to produce inflated (over-fit) coefficients (Yarkoni & Westfall, 2017). Of course, it is important not to over-penalize the coefficients: as a simple rule, the optimal penalization is one that maximizes a model's performance in cross-validation. Using regularized regression approaches combined with cross-validation, prediction models can be based on from a few to tens or hundreds (in fact, even thousands) of inter-correlated predictors (domains, facets or items).

The Present Study

In order to illustrate the ideas discussed above, we employed a large British adult sample ($N \sim 8{,}700$) who had completed a 50-item Big Five personality questionnaire (Goldberg, 1999). Specifically, we predicted 40 outcomes reflecting a variety of life domains from the Big Five personality domains and then from the 50 items used to define the domain scores. Using such a wide range of outcomes yielded a generalizable pattern of findings but it also allowed us to investigate systematic variations among outcomes in their predictability. Specifically, we examined whether differences between domains and items in the prediction of outcomes would track with the breadth of these outcomes as rated by independent judges. The use of a large sample allowed for training models with tens of predictors and validating them in independent and yet sufficiently large subsamples. However, the use of a personality measure with only 50 items, many of which were similar or almost entirely overlapping in content (e.g., "I seldom feel blue" and "I often feel blue")[1], meant that the prediction models could sample markers for only a limited set of nuances from the yet-unknown population of all possible outcome-relevant nuances. Therefore, we had to consider the extent to which items would out-predict domains the *lower-bound estimate* of such a tendency. Most associations were longitudinal over about five years, which somewhat reduced the risk that personality ratings were "contaminated" by outcomes, or the other way around.

Materials and Methods

Participants

This project used data from the National Child Development Study (NCDS), an ongoing longitudinal study of 17,634 individuals born in a specific week in March 1958 in Great Britain, and of a further 929 individuals born in the same week abroad who immigrated to Great Britain before age 16 (Plewis, Calderwood, Hawkes & Nathan, 2004). To date, a wide range of variables reflecting different aspects of the cohort members' lives have been measured in nine separate sweeps, at ages 7, 11, 16, 23, 33, 41/42, 46/47, 50/51 and 55, respectively. Here, data collected in Sweep 8 (2008/2009, age 50/51) and Sweep 9 (2013/2014, age 55) were used (University of London Centre for Longitudinal Studies, 2012, 2014, 2015). Data in Sweep 8 were collected in a 55-minute face-to-face interview, as well as using a self-completion questionnaire posted to participants prior to the interview (Brown, Elliott, Hancock, Shepherd & Dodgeon, 2012). Data in Sweep 9 were collected by first inviting cohort members to participate online and subsequently contacting non-respondents via telephone (Brown & Hancock, 2015).

Measures

Predictors. Personality data were collected as part of the Sweep 8 self-completion questionnaire, using the 50 items from the International Personality Item Pool (IPIP; Goldberg, 1999), measuring Goldberg's (1992) markers for the Big Five (10 items for each trait). The items were answered on a 5-point Likert-type rating scale from *very inaccurate* to *very accurate*. Of the 9,790 respondents in Sweep 8, self-completion questionnaire data were available for 8,787. However, only the 8,719 participants (4,519 female) who had completed more than 80 per cent of IPIP items were included in this study. After median-replacement of 1,326 missing item responses, individuals' item scores pertaining to each domain were averaged to calculate scores on the Big Five personality traits. There were no extreme cases of multicollinearity ($r > .80$)

1 http://ipip.ori.org/newBigFive5broadKey.htm

between any items. Correlations between the Big Five domains ranged from .06 to .38 in magnitude (median = .24).

Outcome Measures. Variables collected in Sweep 8 and Sweep 9 were screened as possible outcome candidates. For 7,621 (3,999 female) of the 8,719 participants in this study, data from both Sweep 8 and Sweep 9 were available. For the remaining 1,098 (520 female) individuals, data were only available from Sweep 8. Most outcome candidates were selected based on previous literature and a theoretical rationale that they could be related to personality. A few outcomes, such as 'Partner's Age', 'Attending Concerts or Theatre' and 'Eating Out' were included for purely exploratory purposes. During the outcome selection process, outcomes which were answered on an ordinal, interval or continuous scale were given preference over binary ones. Similarly, where available, preference was given to outcomes measured in Sweep 9 over those measured in Sweep 8 to avoid criterion contamination, whereby an outcome measured concurrently with personality may have affected personality ratings or the other way around. For instance, how often one currently sees friends may influence extraversion ratings. Only five of the selected outcomes were based on data collected in Sweep 8. All 40 outcomes are shown in Table S1 of the Supplementary Material (available from https://osf.io/2efnr), alongside the names of the original NCDS variables and a record of any changes made to these variables. The variables can also be seen in Table 1 of the main text. For detailed descriptions of how NCDS variables were measured, including exact wordings of interview or questionnaire items, see the NCDS documentation (Brown et al., 2012; Brown & Hancock, 2015).

Thirteen of the outcomes were derived from the original NCDS variables, either by creating binary variables from categorical variables with more than two levels, or by combining two NCDS variables supposedly measuring the same outcome into a single variable (see Table S1). The remaining 27 outcomes were directly based on a closely corresponding NCDS variable. However, in many cases, minor changes were made to the variables (see Table S1). For 16 variables, the original coding of the NCDS variable was reversed in order to make the interpretation of personality-outcome relations more intuitive. For instance, the coding of the outcome 'Volunteering' that originally ranged from (1) *at least once a week* to (4) *never* was reversed in order for higher values to correspond to volunteering more frequently. Following screening for outliers and normality, extreme values for a few cases were either excluded or capped at a certain value for the outcomes 'Income', 'Body Mass Index' and 'Working Hours per Week'. In addition, three variables, 'Alcohol Units per Week', 'Income' and 'Body Mass Index', were log10-transformed to make their distributions more normal. Table S1 also provides information on the outcomes' measurement levels, and the coding of response options for binary and ordinal outcomes after any changes had been made. In total, 13 outcomes were binary, and modeled using logistic regression. The remaining 27 outcomes had either been measured on interval or continuous scales (7 outcomes) or on ordinal scales with at least four levels (20 outcomes); 23 of these variables were fairly normally distributed and modelled using the Gaussian distribution, whereas four variables—'Sport (R)', 'Volunteering', 'Internet Use (R)' and 'Number of Children'—were heavily skewed towards smaller values and were thus treated as having a Poisson distribution when building outcome prediction models. This is the reason why for 'Sport (R)' and 'Internet Use (R)' the original NCDS coding was not reversed, meaning that higher variable values correspond to using the internet less and doing less sport (the label (R) is included to remind the reader of this counterintuitive coding).

Outcome breadth *versus* specificity. In order to be able to test the hypothesis that items tend to confer more incremental predictive value for relatively narrower outcomes, we asked 19 individuals (aged between 19 and 30 years; 12 had no formal connection with psychology and the rest were either former or current students or had otherwise some experience of psychological research) to rate each of the 40 outcomes in terms of their breadth *versus* (behavioral) specificity. The raters were presented with the following instruction: "Below is a list of variables that researchers may use to characterize individual differences in behavior, attitudes, socioeconomic performance, health, and so forth. To what extent does each of these variables represent a relatively specific type of behavior? At the other end of the spectrum, the variables may represent broad life outcomes, reflecting the contributions of multiple types of specific behaviors that could happen over longer periods of time. Please rate each variable on this dimension of specific behaviors *versus* broad life outcomes." The outcomes were rated on a 5-point Likert scale, with endpoints marked as *Specific behavior* (1) and *Broad life outcome* (5). Across the 40 outcomes and 19 raters, the intraclass correlation (*ICC*) of single raters was .18, whereas the *ICC* of average ratings was .81 (calculated using the *psych* package; Revelle, 2017). For each outcome, we averaged the ratings of the 19 raters.

Statistical Analyses

The selection of NCDS variables as outcome candidates, outcome creation from these via recoding, initial checks for outliers and normality as well as deletion of extreme values and log10-transformations as described above were carried out in SPSS, version 21. All subsequent analyses were carried out using R 3.4.2 (R Development Core Team, 2016).

Model Building and Outcome Prediction. In order to avoid over-estimating models' predictive strength due to over-fitting, the sample was split into a training sample (3/4) and a validation sample (1/4). We chose to use a notably larger proportion of participants for model training, because the larger the training samples, the more precise the models tend to be and thereby the higher predictive accuracy in the validation samples they generally allow for (for a parallel in GWAS and polygenic scoring, see Cesarini & Visscher, 2017). Prediction models were fitted in the training sample, from which regression weights were obtained, and then applied in the validation sample for outcome prediction. Squared correlations between predicted values and observed values were used as estimates of model prediction strength. The process of model training and outcome prediction was repeated one hundred times in random splits of the sample.

Prediction models for each outcome were built using a penalized regression, which shrinks regression coefficients towards zero by imposing a penalty on their combined size. Penalized regression can be employed when the number of potential predictors in a model is high, such as in GWAS (Waldmann, Mészáros, Gredler, Fuerst & Sölkner, 2013), and when the predictors have high inter-correlations. Two widely-used forms of penalized regression are ridge (Hoerl & Kennard, 1970) and LASSO (Tibshirani, 1996). Ridge regression applies a penalty to regression coefficients that depends on the sum of the *squares* of these coefficients, whereas the LASSO penalty depends on the sum of the *absolute values* of these coefficients. Both approaches have limitations. Ridge regression tends not to increase model parsimony as it rarely shrinks regression coefficients to zero. In contrast, LASSO leads to sparser (with more zero coefficients) and thereby more parsimonious and readily interpretable models, but it has a downside of often randomly selecting one out of many correlated predictors, setting coefficients for others to zero

(ridge regression tends to shrink coefficients for correlated predictors towards each other; Waldmann et al., 2013; Friedman, Hastie & Tibshirani, 2010). This means that LASSO solutions are not unique. We thus employed the elastic net penalty (Zou & Hastie, 2005), which combines both ridge and LASSO penalties, mitigating limitations associated with each of them alone (e.g., Chapman et al., 2016). Elastic net tends to yield parsimonious models (due to the LASSO penalty), in which groups of correlated predictors are treated in the same way (due to ridge penalty), either all being included or excluded from the model (Waldmann et al., 2013; Zou & Hastie, 2005). We modeled binary outcomes using the binomial link (logistic regression), outcomes with a Poisson distribution using the Poisson link and the remainder of outcomes using the Gaussian link. The optimal regularization parameter lambda (λ) was obtained using 10-fold cross-validation within the training sample such that it minimized cross-validated error across the folds[2]. These procedures were carried out by the *glmnet* package for R (Friedman et al., 2010).

Adjustment. We did not control for age as all participants were born in the same week, but we adjusted for gender by including it as a covariate into the prediction models built in the training samples (we did not residualise the outcomes for gender outright because not all of them were continuous). When obtaining predicted values in the validation samples, the regression weight of the gender covariate was set to zero, in order to obtain predictions only from personality while having regression coefficients for personality variables that controlled for potentially spurious sex effects. This was done to avoid inflating what would be interpreted as personality's predictive share in the outcomes.

All R scripts are publicly available at http://osf.io/z9pr2. NCDS data are available from the UK Data Service, University of Essex (http://www.ukdataservice.ac.uk).

Results

Model Calibration

We carried out a simulation to test the degree to which our procedure of training models in one partition of the sample using the elastic net penalty and validating them in another partition would guard against more complex (item-based) models out-predicting more economical (domain-based) models for purely statistical reasons (e.g., over-fitting). We simulated data similar to our empirical data (N = 9,000; 50 items grouped into five traits; 40 outcomes), but with a particular underlying structure: five latent traits [$N(\mu = 0, \sigma^2 = 1)$] contributed to 50 observed variables [10 for each latent trait; $N(\mu = 0, \sigma^2 = 1)$] and any number between one and five of them could also contributed towards the outcome variables [$N(\mu = 0, \sigma^2 = 1)$]. The factor loadings of items on their latent traits were in a realistic range, varying from about just under .40 to just over .70 (they ranged from .35 to .76 in our real data), and the contributions of latent traits towards the outcome varied from close to zero to potentially about .60 (mostly < .20, very rarely > .40). We then applied the same procedure on these simulated data that would be applied on the empirical data (comparing the predictive accuracy of "items" and the five "domain" scores they would make up), with a focus on the degrees to which item- and domain-based models predicted

---

[2] 10-fold cross-validation means that the training subsample was divided into 10 yet smaller equal subsamples and the model was cross-validated once in each fold, while training it in the remaining nine folds, and the coefficients were averaged across all runs. The regularization parameter λ was chosen (from a range of possible parameter values) such that it produced the best average prediction in the cross-validation folds. The procedure is implemented in the *glmnet.cv* function of the *glmnet* package.

the outcomes. We repeated the procedure 100 times. The simulation R-code is available at https://osf.io/2fnqm.

In most cases (83%), domain-models out-predicted item-models (as per underlying data-generating model), with the average predictive accuracy of item-based models being about 96% of the accuracy of domain-models (1% and 99% quantiles of the ratio of item-model predictive power to domain-model predictive power were 0.80 and 1.02, respectively). Therefore, if personality-outcome associations were indeed driven by the purported latent traits that the item composites were designed to measure, our statistical procedure was likely to correctly show that domains out-predicted items. By implication, this meant that when item-models would out-predict domain-models in real data, it would tend to correctly indicate that the associations were at least in part driven by the unique characteristics that the items reflect. Thus, the procedures employed in this study were likely to effectively guard against more complex models out-predicting more economical ones for artefactual reasons.

Predictive Strength of Item- and Domain-Models

On average, approximately two thirds of IPIP items and four to five Big Five domains were included in item- and domain-models respectively (i.e., had non-zero elastic net regression coefficients, which are given the Table S2 of the Online Supplementary Material; https://osf.io/8sjyr). Table 1 shows the mean variance explained (R^2) in each outcome by item- and domain-models across 100 replications; we also report the 99% confidence intervals (CI) for these average estimates, calculated based on their variance across the 100 replications.

Domain-models for 33 outcomes and item-models for 35 out of the 40 outcomes predicted more than 1% of variance ($R^2 > .01$). Mean variance explained by domain- and item-models varied considerably across the outcomes, and was highest for outcomes such as 'Sleeping Enough', 'NVQ (Education)', 'Feeling in Control', 'Racist Attitudes', 'Optimism' and 'General Health' ($R^2 > .10$). Among the outcomes, the least predictable from personality characteristics were 'Frequency of Exercising', 'Partner's Age', 'Divorced', 'Stopped Smoking', 'Diabetes', 'Number of Children' and 'High Blood Pressure' ($R^2 < .01$ for domains). On average, across all outcomes, item-models explained 5.45% of variance, while domain-models explained 4.18%, (medians were lower, 3.36% and 2.46%, respectively, indicating a positive skew). In other words, the average item-level prediction exceeded that of domain-level prediction by about 30%.

For 37 outcomes, item-models tended to out-predict domain-models; for 33 of these, the 99% CIs of average item- and domain-level predictions did not overlap. For three outcomes ('Cigarettes per Day', 'High Blood Pressure' and 'Partner's Age') the predictions were roughly similar in magnitude (difference in $R^2 < .001$). Although these outcomes were among the least predictable from personality characteristics in the first place, generally the *ratio* of the strengths of item-model and domain-model predictions was not significantly linked with prediction strength of either domain- or item-models, suggesting that, as a general tendency, items were comparatively stronger predictors for outcomes regardless of the overall degree to which these could be predicted from personality.

When only considering the 33 outcomes for which domain-models on average predicted more than 1% of variance ($R^2 > .01$) in order to avoid including inflated estimates where the outcome was not predicted very much at all, the prediction improvement of item- over domain-

models ranged from 3.58% ('Cigarettes per Day') to 94.75% ('Voted in Last Election'). For these 33 outcomes, a paired Wilcoxon signed rank test showed the difference in variance explained between item- and domain-models to be significantly different from zero ($p = 2.33 * 10^{-10}$). These results stand in stark contrast with the simulation results presented above, suggesting that personality trait-outcome associations cannot be fully accounted for by the (underlying) Big Five domains.

Table 1. *Explained variance (R^2) from item- and domain-models.*

	Domain-models			Item-models			Δ	%	*N*
	Mean	99% CIs		Mean	99% CIs				
Sleeping Enough	.145	.142	.149	.158	.155	.161	.013	9	2175
NVQ (Education)	.136	.133	.140	.212	.208	.216	.076	56	1905
Feeling in Control	.107	.104	.110	.123	.119	.126	.016	15	1882
Racist Attitudes	.106	.104	.109	.118	.115	.121	.012	11	2168
Optimism	.104	.101	.107	.119	.115	.122	.015	14	1875
General Health	.100	.097	.103	.116	.113	.119	.015	15	1891
Depression	.099	.096	.102	.118	.115	.121	.019	19	1887
Income	.076	.072	.079	.115	.111	.119	.039	52	1033
Social Class	.071	.069	.074	.112	.108	.115	.040	56	1497
Internet use (R)	.063	.061	.065	.091	.089	.094	.029	45	1888
Relationship Satisfaction	.055	.052	.057	.065	.062	.067	.010	18	1837
Seeing Friends	.050	.048	.052	.057	.055	.060	.007	15	2174
Managing Financially	.050	.048	.053	.071	.068	.073	.020	40	1886
Attending Concerts or Theater	.048	.046	.050	.063	.060	.065	.014	29	1887

	Domain-models			Item-models			Δ	%	*N*
Volunteering	.046	.044	.048	.057	.055	.059	.011	25	1886
Eating Out	.036	.035	.038	.041	.039	.042	.004	12	1887
Part-time Employed	.028	.026	.030	.035	.033	.036	.007	23	1539
Job Satisfaction	.028	.026	.030	.031	.029	.033	.003	12	1494
Frequency of Alcohol Use	.026	.024	.028	.045	.042	.047	.018	70	1756
Practicing Religion	.025	.023	.026	.031	.030	.033	.006	26	2170
Alcohol Units per Week	.024	.022	.026	.035	.033	.038	.011	45	1265
Sport (R)	.024	.022	.025	.026	.024	.027	.002	9	1886
Body Mass Index	.023	.021	.024	.037	.036	.039	.015	65	1783
Voted Conservative vs. Labour	.022	.020	.024	.031	.028	.033	.009	41	911
Number of Cars	.021	.019	.023	.032	.031	.034	.012	55	1902
Voted Liberal	.019	.018	.021	.025	.023	.027	.006	30	1210
Receiving Benefits	.018	.017	.020	.022	.020	.023	.004	19	1880
Working Hours per Week	.018	.016	.019	.030	.028	.032	.013	72	1500
Never Smoked	.016	.015	.018	.028	.027	.030	.012	73	1888
Never Married	.016	.015	.017	.018	.017	.019	.002	13	1904
Voted in Last Election	.016	.015	.018	.032	.030	.034	.016	95	1873
Cigarettes per Day	.013	.010	.016	.013	.010	.016	.000	4	249
Self-employed	.011	.010	.012	.019	.018	.021	.008	74	1539
High Blood Pressure	.009	.008	.009	.009	.008	.009	.000	0	1889
Number of Children	.008	.007	.009	.015	.013	.016	.006	78	1905
Diabetes	.004	.004	.005	.005	.005	.006	.001	23	1889

	Domain-models			Item-models			Δ	%	*N*
Stopped Smoking	.004	.003	.005	.008	.007	.010	.005	114	918
Divorced	.003	.003	.004	.011	.010	.012	.008	240	1648
Partner's Age	.002	.002	.003	.002	.001	.002	-.001	-22	1518
Frequency of Exercising	.001	.000	.001	.004	.003	.004	.003	484	1687

Note. CI = Confidence Interval; Δ = Difference in the average R^2 of item- and domain-models (positive values show the incremental value of item-models). % = Difference in the average R^2 of item- and domain-models in percentage metric. *N* = number of participants in the validation sample (i.e., training sample size is three times *N*).

The Breadth of Outcomes

We correlated the outcomes' average breadth ratings, given by the 19 raters, to the degrees to which they were predicted by either domains or items, and the difference and ratio between the two kinds of predictions (*r*-to-*z* transformed columns 2 and 5 of Table 1 or the ratio between them). None of the correlations was sizeable (Spearman's *rho* = -.09 to 0). For example, among the outcomes mostly strongly out-predicted by the item-models were specific behaviors referring to attending cultural events, using the internet or consuming alcohol as well as broad outcomes, such as educational qualification, income or BMI. This suggests that the variability among the 40 outcomes in their behavioral specificity *versus* breadth had little to do with how well they could be predicted from either items or domains, or with the degree to which items conferred incremental predictive value over domains. Items tended to out-predict domains regardless of the breadth of what was predicted.

Domain-Level Predictions Were at Least in Part Driven by Nuances

The typical predictive advantage of item-models over domain-models was arguably only moderate (about 30%). But it is important to realize that this *observed* advantage was unlikely to have entirely accurately revealed the degree to which the nuances captured by the items tended to predict the outcomes, on top of or rather than the latent traits purported to underlie the domains scores. This is because the predictive value of nuances was always *included* in the domain scores, likely inflating the predictive value of the domain scores compared to what it would have been without these nuances included. In other words, the Big Five domains *per se*, independently of any particular items that happened to be included in their operationalization (but could *not* have been included, had the test constructors chosen alternative items that were equally reflective of the domains but with different unique outcome-associations), could have done worse in the prediction. This reasoning is of course based on the assumption that the underlying traits that the Big Five scales are supposed to measure exist independently of how particular questionnaires approximate them—but this is a *de facto* standard assumption in personality research anyway (Mõttus, 2016).

In order to address this possibility, the predictive power of item-models should be compared to domains operationalized *independently of these particular items* (e.g., by using another questionnaire with items that only measure the domains and not nuances). We did not

have such data and we doubt that anyone has. However, as a *post-hoc* analysis, we could remove a few items that most strongly predicted each outcome from the Big Five scales and re-estimate their predictive power after this (R Code available from https://osf.io/bce2h). Assuming that the underlying traits that the domain scores were designed to approximate indeed exist independently of the particular items aggregated into them (Mõttus, 2016), the reduced scores should have measured the same underlying traits as full scores, possibly barring a small drop in measurement reliability. Therefore, if the associations were driven by the purported latent traits, their predictive value should have dropped minimally when a few items were removed[3].

For each outcome, therefore, we compared the predictive power of the model based on 50 items to that of five Big Five domain-models, with domain scores calculated based on either 49, 48, 47, 46 or 45 items in total; that is, one to five of the most predictive items were removed from domain scores, regardless of which domain they fell into. As above, the models were trained and validated in independent samples and the procedure was repeated 100 times in random sample splits for training and validation, respectively. Dropping only the most predictive item (identified using elastic net regression, as above) from the domain score this item initially happened to belong to reduced the average (across the 40 outcomes and 100 permutations for each) predictive power of domain-models by about 6%, whereas also removing the second, third, fourth and fifth most predictive item from the domain scores these items happened to belong to reduced the average predictive power of domain-models by about 11%, 14%, 16% and 19%, respectively. In most cases, removing the five most predictive items left the shortest scale (i.e., scale from which the most items had been removed) with eight items instead of ten, but for five outcomes up to three and for two outcomes up to four items out of ten were removed from what would become the shortest scale. For reference, when we removed five randomly chosen items from among the 50 items making up the five domains, then average amounts of variance accounted for by domain-models only decreased by about 1%.

Assuming that domain-models with the five most predictive items removed from the domains (i.e., leaving them based on 45 items instead of 50) constituted at least a somewhat fairer comparison for item-models than domain-models with the most predictive items included in their variance (because the scores still measured the same domains, regardless of whether they contained eight or ten items), the average predictive power of item-models (R^2 = .0545) was about 61% higher than the average predictive power of the domain-models based on fewer items (R^2 = .0335). Of course, this could also be an underestimate, because other (than the "top-five") nuances uniquely predictive of some outcomes were still included in the domain scores and potentially still inflated the estimated predictive power of domains *per se*.

Specifically, Figure 1 shows the average predictive value of both domain- and item-models when up to 10 (i.e., 20%) most predictive items were removed from them (for 30 of 40 outcomes, the shortest scale retained six or more items, for eight outcomes the shortest scale retained five items and for two outcomes only three or four items were retained in the shortest

3 Assuming a reliability of .80 for a 10-item scale, dropping one, two, three or four items from it would be expected to reduce its reliability to .78, .76, .74, and .71, respectively, according to the Spearman-Brown formula. Then, assuming a true correlation of .20 with a perfectly measured outcome bounded by the above-listed reliability coefficients on the personality scales side, removing one, two, three or four items would reduce the correlation by 1%, 2%, 4% or 6%, respectively. Note, however, that most predictions were driven by multiple scales, diluting the effect of dropping one or a few items from one or some scales.

scale; note, however, that outcomes were mostly predicted by four or five scales and most scales contained more items than the shortest scale). In the figure, the predictive values are grouped into quartiles according to the degree to which models based on 50 items predicted the outcomes (i.e., the top quartile represents the average predictive value of the 10 outcomes most accurately predicted from personality characteristics, which roughly corresponds to the top 10 rows of Table 1). Both item- and domain-models tended to loose some of their predictive power as ever more items were removed, and this tendency was fairly similar regardless of the degree to which the outcomes had been predicted in the first place (across the four quartiles, removing ten items decreased the average predictive value of domain-models by 26% to 34%; the average across all 40 outcomes was 31%). However, while the predictive advantage of item-level models did not pertain to a few "top nuances" for the most predictable outcomes (in the top quartile, the predictive advantage of items was 27% with all items included and it was still 21% with ten top items excluded), in outcomes where the overall amount of variance accounted for by personality was smaller this seemed to be the case. For example, for the least predictable outcomes, removing only a few of the most predictive items resulted in domain-models being on par or even outperforming item-models. This tendency suggests that the more predictable from personality characteristics an outcome was, the more this prediction was driven by nuances. These findings indicate that domain-level predictions are often likely to be at least in part driven by the nuances that happen to be included in them.

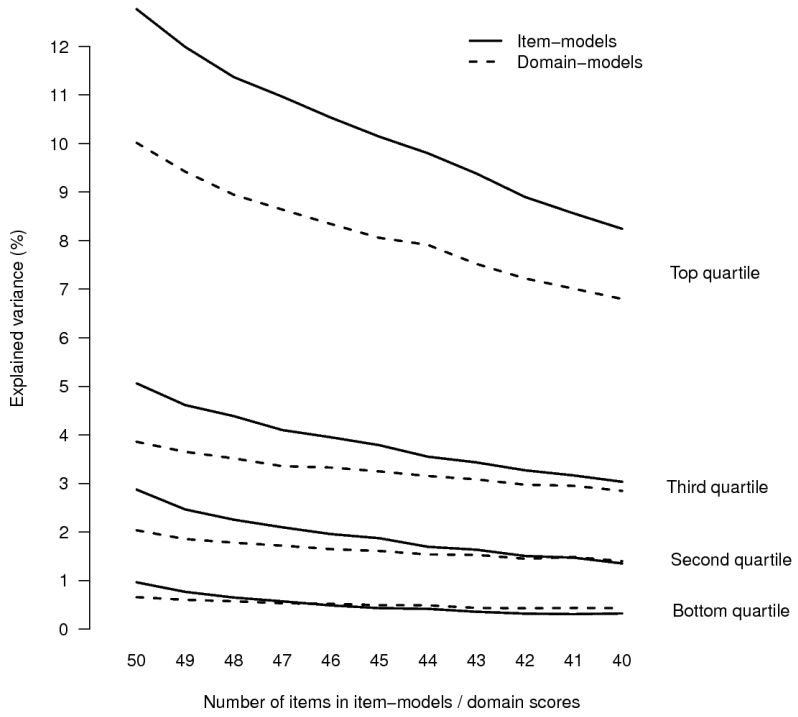

Figure 1. The predictive power of item- and domain-models based on 50 to 40 items (with most predictive items iteratively removed). Outcomes are grouped based on the degree to which they were predicted by 50-item-models.

Item-Level Predictions Were Mostly Not Driven by Domains

In order to estimate the extent to which item-level predictions were driven by their common variance, ostensibly reflecting underlying domains, we carried out another *post-hoc* analysis. Specifically, we re-ran the item-level predictions (as above, training models on 75% of the sample and validating the models in the remainder of the samples, and repeating this procedure 100 times) using the *residual* variance of items (Mõttus, Kandler et al., 2017; Mõttus et al., under review). Specifically, items were conditioned on the scores of all five domain scores using linear regression, and the residuals were saved (the item being residualized was removed from its intended domain score at the time to avoid regressing the item on itself). The item residuals tended to correlate highly with item scores before residualizing, with r = .69 to .94 (M = .83).

On average, the models based on item residuals explained 5.11% of variance in the 40 outcomes, and for 33 of them the percentage exceeded 1%. Therefore, removing the domain-variance from items only attenuated their average predictive value by about 6%, and the average item residual-model still out-predicted the average domain-model by about 22%. This result reinforces the notion that much of the predictive value of personality for outcomes tends to stem from the characteristics that are aggregated into traits rather than whatever underlying causal entities the aggregates (items' common variance) are purported to approximate.

Discussion

We explored the usefulness of personality questionnaire items as personality markers (in a way analogous to genetic markers, SNPs) for representing personality-outcome associations. We found that prediction models based on questionnaire items (similar to prediction models built from GWAS data in genetics) accounted for a non-zero share of variance in most outcomes. In fact, the item-models mostly showed greater prediction strength than models built from Big Five domains, with an average of 30% more variance explained. We also made a case that this percentage likely underestimated the degree to which the unique variance in items was predictive of outcomes on top of, or even rather than, the traits purportedly underlying the Big Five domains. This is because the domain-level predictions were inflated by the unique variance of the items included in them, and items that had been residualized for the domains predicted the outcomes only slightly worse than items that included the domain variance. Furthermore, we found that the degree of the incremental predictive value of items over domains did not depend on the breadth (*versus* behavior specificity) of the outcome.

These findings were based on a large sample of more than 8,700 participants, mostly longitudinal associations, and a wide range of diverse outcomes. Moreover, we used rigorous statistical procedures that guarded against model over-fitting, as shown by the simulation. Importantly, unlike most studies that link psychological constructs with outcomes by fitting statistical models and estimating the performance of these models in the same sample(s), in this study outcomes were predicted from models that had been trained in independent people. This means that we quantified genuine *predictive power* rather than just correlations. Collectively, the findings suggest that particular personality-outcome links often pertain to specific personality characteristics rather than the broad Big Five traits these characteristics are ostensibly only indicators of.

Personality Links Are Pervasive, Although Often Expectedly Weak in Magnitude

The present findings reinforce the conclusion that the associations of personality characteristics with life outcomes are ubiquitous (Ozer & Benet-Martínez, 2006; Roberts et al., 2007). Even though the selection of outcomes was not entirely random in this study, we can hypothesize that most markers of individuals' socioeconomic success, health or social behavior can, to some extent, be predicted from personality characteristics included in omnibus personality models. Of course, this does not mean that the links are necessarily reflective of causal contributions of personality to these outcomes (Mõttus, 2016; Mõttus, Marioni, & Deary, 2017), but sometimes they may be—patterns of behavior can have consequences.

However, the effect sizes were generally weak, with on average slightly over 4% of variance being predicted from the five Big Five scores and about 5.5% from their 50 items. This has several mutually non-exclusive explanations. One possibility is that the tendency for low effect sizes is an accurate reflection of reality: anything that people differ in is likely to have a myriad of causes—many of them possibly idiosyncratic—and personality differences at any single point may constitute only a fraction of them. Another possibility is that the 50-item measure used in this study covered only a small sample of potentially relevant personality characteristics, either as domains or nuances. It is only too likely that the prediction models—especially item-level models—could have performed much better had a more comprehensive measure been used; we'll also discuss this issue below. A third explanation for the relatively modest effect sizes is that they resulted from out-sample predictions and were little, if at all, upwardly biased by over-fitting—unlike most results reported in literature (Yarkoni & Westfall, 2017). And yet it should be noted that nearly a quarter of outcomes did share over 10% of their variance with 50 item-level personality characteristics.

What Do the Findings Tell Us About How Personality Intersects With Outcomes?

We expected personality characteristics, especially the broad Big Five traits, to be comparatively stronger predictors of broad life outcomes that aggregate the accumulating consequences of a wide range of behaviors, thoughts, feelings and aspirations (Asendorpf et al., 2016; Wittmann, 1988), as opposed to more specific behavioral outcomes. This would have been consistent with the possibility of there being broadly acting underlying personality traits casting their non-specific influences through narrower characteristics (nuances) whose (unique) relevancy only depend on specific outcomes. Although the nuances would then be etiologically more proximal to the outcomes and might therefore have stronger links with them, it is unlikely that personality questionnaires, especially short inventories such as the 50-item IPIP, cover the nuances specifically relevant for each and every outcome. For the most part, then, the nuances reflected in test items would serve as mere indicators (measurement devices) of the broader underlying traits. But this did not appear to be the case: the degree to which outcomes were predicted from personality did not track their breadth/specificity and items tended to out-predict domains for broad and specific outcomes alike.

But could it be that the items' incremental predictive validity results from item-outcome overlap (Mõttus, 2016; Mõttus, Marioni, & Deary, 2017)? This was unlikely for the present findings. For example, among the outcomes for which items made the biggest incremental contribution (in relative increase terms) were voting, being self-employed, time spent working, smoking, alcohol use, BMI, educational qualification, occupational social class, number of cars

owned, income and internet use but none of the IPIP items made any reference to them. Indeed, for most outcomes, it is hard to see how they would be more obviously connected with any individual IPIP item than with any Big Five domain. And yet items were collectively more strongly linked with the outcomes than the domains, even if the domain-related variance had been removed from the items. Moreover, the more outcomes were incrementally predicted by items, the less likely this was driven by only a few items. For example, for the ten most predictable outcomes, even after dropping the ten most predictive items item-models tended to out-predict domain-models by nearly the same ratio than the models with all items included.

A plausible interpretation of these results is that the associations of personality and outcomes *do not* generally pertain to the ostensible underlying traits. Instead, outcomes may be highly polynuanced—linked with a wide range of specific personality characteristics—such as phenotypes are generally polygenic (Chabris et al., 2015). If so, personality trait scores are correlated with outcomes because questionnaire items sample from among the nuances that are either *directly* linked with the outcomes themselves or are linked with other nuances that are relevant for the outcomes. The latter possibility of *indirect* associations between items and outcomes contributing to predictive power recasts the idea that genetic markers (SNPs) can be linked with phenotypes not only because they represent genetic variants directly relevant for the phenotype but also because they are in linkage disequilibrium with the directly relevant variants (i.e., serve as proxies). In the personality context, such "linkage" may arise from direct causal associations among the nuances or their links with overlapping motivational characteristics (Cramer et al., 2012; Wood, Gardner & Harms, 2015), among other reasons.

To the extent that this scenario applies, because personality test items constitute samples of markers of potentially outcome-relevant nuances, aggregating them into broad traits almost inevitably filters out some of the outcome-relevant information—which is exactly what the present results tended to show. The present findings do not only point to where in the (descriptive) personality trait hierarchy the outcome-relevant ingredients may lay, but also reinforce the concept of nuances as potentially useful descriptive—and maybe even explanatory —units of personality (McCrae, 2015; Mõttus, Kandler et al., 2017; Mõttus et al., under review).

But Why Do We Aggregate in the First Place?

Specific personality characteristics (tendencies for specific behaviors, feelings, cognitions and motivations), as reflected in single items, are typically aggregated into composite scales in order to increase the reliability of measurements, allow for more parsimonious models and because it is hoped that the aggregates reflect some underlying, etiologically homogeneous and causally potent properties of human mind. What has been proposed in this article may not seem in lockstep with these aspirations. So?

Indeed, the ratio of measurement error to substantive (non-error) variance is larger in single items than in aggregate trait scores, which could limit the value of item-based analyses. However, this is primarily a problem for studies based on small samples where model parameter estimates are less stable and aggregation of observations *per person* helps to increase their reliability. In sufficiently large samples, such as the one used in this study, parameter estimates are more stable even with lower reliability of single measurements, as the aggregation of observations *across persons* compensates the low reliability of single measurements (Goldberg, 1993). But even with smaller samples ($N \sim 1{,}000$) than used here, parameter estimates pertaining

to single items (e.g., age or gender differences, or associations with outcomes such as BMI) tend to be consistent across studies and must therefore be reasonably reliable (Mõttus et al., under review).

Likewise, item-based models for predicting outcomes are apparently less parsimonious than those based on higher-order traits, simply because there are more items than their aggregates. Generally, science strives for simplicity and parsimony, *ceteris paribus*. But it is exactly this latter clause—all else being equal—that is important here. First, if and when items do allow for outcome predictions that are more accurate, then it might appear that reliance on what appears as less parsimonious at face value has some benefits. It has been argued that it is exactly prediction that psychology should strive for, rather than grossly simplified explanatory models with commensurately low practical value for describing what is going on in the real world (Yarkoni & Westfall, 2017). Second, relying on composites for, say, causal explanations requires the composites to have an appropriate ontology: they need to reflect something real about individuals rather than just being summaries of psychological 'stuff' (Mõttus, 2016). If the composites were just convenient summaries of items, it would still be the items that have to carry the explanatory weight in the end and we might just as well represent the associations using these (cf. Wood, Gardener, Harms, 2015). Doing so would not mean doing away with personality as all the information pertaining to composite traits—and some more—would be retained, even though modeled differently.

Again, we may rely on a parallel with genetics. The realization that the genetic architecture of complex phenotypes is so complex that it does not lend itself for *a priori* hypothesizing is exactly what has finally allowed geneticists to predict phenotypic variance from genome-based observations (see Plomin & von Stumm, 2018). That is, it is exactly suppressing the strive for apparent parsimony that has been useful—because reality cannot always be represented parsimoniously. In a way, of course, realizing and accepting that things are complicated (e.g., polygenic or polynuanced) can allow for the emergence of new, higher-order principles that focus less on which specific elements of a complicated system are inter-linked but look for some general organizational principles of the system.

It could be argued that domain-level predictions are particularly useful because they allow for generalized explanation. For example, an observed correlation between Conscientiousness and longevity can be explained by a variety of behaviors that conceptually fall under this domain (e.g., being mindful about one's health and able to resist urges to behave in unhealthy ways, adhering to medical advice and treatment), regardless of whether these have been directly captured in particular questionnaires. However, to the extent that personality-outcome associations are actually not driven by domains such as Conscientiousness but the specific characteristics that happen to be captured by the questionnaires or somehow in "linkage" with them, such generalizations may be particularly dangerous—they would need to be tested rather than assumed. Knowing the patterns of correlations (or "linkage") among these characteristics (the basis for domains), can guide our hypothesizing as to which characteristics could be relevant in addition to those that have been directly measured and linked with any given outcome. But we do not necessarily need domains *per se* to explain the associations.

But it is also important to realize that identifying item-level (or nuance-level, by inference) associations does not preclude aggregation. For example, outcomes could then be predicted from polynuance scores, which are weighted aggregates of items (weights being the associations of the

items with the outcomes), exactly as phenotypic variance is being predicted from polygenic scores, which are weighted aggregates of genetic markers (Plomin & von Stumm, 2018). In addition to the prediction of outcomes for which the polynuance scores were initially created, other outcomes could be predicted, or associations between polynuance scores created for different outcomes could be calculated. This would allow exploring the extents to which different outcomes either correspond to different personality profiles or are independent with respect to their personality-related mechanisms. Again, this recasts an extremely useful concept in genetics, genetic correlation (e.g., Neale & Maes, 1996), which quantifies the extent to which different phenotypes are linked with overlapping genetic variance. Geneticists study patterns of genetic correlations among phenotypes to learn about their genetic etiology (e.g., Bulik-Sullivan et al., 2015; Plomin & von Stumm, 2018), and personality psychologists could study patterns of "personality correlations" (e.g., correlations between polynuances scores for different outcomes) to learn more about how personality relates to outcomes. These patterns may inform us on some of the general principles regarding how personality intersects with phenomena outside the personality domain.

What Do the Findings Tell Us About Personality Traits?

As recently outlined by Baumert and colleagues (2017), one of the most important questions regarding the etiology of personality traits—defined as correlated patterns of behavior, thinking and feeling—is whether each of them corresponds to a shared set of processes (amounting to a latent common cause) exclusive to this particular trait ('correspondence') or whether they arise from more complex interaction processes among some more basic components of personality ('emergence'). It is possible to think that these basic components represent what we have termed as personality nuances. To the extent that the former scenario applies and the Big Five domains approximate the shared processes constituting latent causes of particular traits, one could expect personality-outcome associations to be mostly driven by the domains. To the extent that behavior, thinking and feeling coalesce into what appear as traits due to more widespread interactions among them, there is less reason to think that the personality-outcome association should be driven by the traits *per se*, because the behaviors, thoughts and feelings that give rise to them are causally autonomous and they are not exclusively aligned with any one trait alone. If so, the present findings may be more in line with the 'emergence' explanation of traits, although not directly supportive of it.

Implications for Behavioral Interventions

Findings that personality traits are linked with a range of positive and negative life outcomes have lead researchers to consider the possibility of intervening on relevant personality traits to obtain desirable changes in these outcomes. For example, Jokela and colleagues (2013) discuss the possible effect of increasing Conscientiousness on improving life-expectancy. When and to the extent that personality-outcome associations are driven by nuances rather than broadly acting underlying trait domains, potential interventions aiming to change outcomes by changing personality ought to first identify the most relevant nuances for these outcomes and then specifically target these. On the one hand, this could be easier than targeting domains, which would be a more attractive course of action if the associations were driven by domains *per se.* For example, changing a habit is likely to be easier than changing the whole collection of behaviors, thoughts, feelings and motivations that Conscientiousness encompasses. On the other hand, if the number of relevant nuances is large, selecting the best targets may be complicated.

Nevertheless, our findings suggest that personality-based interventions may generally be more successful when focusing on more specific behavioral, cognitive, affective and motivational tendencies.

Limitations and Future Directions

Probably the biggest limitation of this study is that the 50-item Big Five questionnaire, IPIP, is likely to cover only a very limited set of nuances. Not only is the sheer number of items small, but these items also tend to overlap in their content, for example 'I seldom feel blue' and 'I often feel blue', or 'I have a vivid imagination' and 'I do not have a good imagination'. As a result, the potential benefits associated with item-model prediction may be much more substantial when more comprehensive personality measures are used. Moreover, future studies should go beyond a pre-conceived and contrived construct space. Specifically, most existing personality questionnaires are explicitly designed to include sets of items that each measure one of the Big Five domains and nothing else. Within these sets, items are selected to maximize their common variance in order to ensure questionnaires' internal consistency—that is, items are designed to overlap in content. Consequently, by design, most personality questionnaires are likely to measure only limited sets of nuances, even when they include large numbers of items. In addition to using a range of *existing* questionnaires, future research should investigate whether outcome prediction can be further enhanced by deliberately selecting a diverse range of items and covering as broad a range of nuances as possible. To do so, more nuances should be identified, for instance, by subjecting data from large personality item pools to clustering procedures (Condon, Roney & Revelle, 2017).

As item-level analyses appear to confer substantial additional predictive value, reliably detecting this will require large samples—another lesson we can learn from genetics. It is a common practice in GWAS studies to aggregate samples, often using harmonized or linked measures of the outcomes (Davies et al., 2015; Hill et al., 2017). Also, GWAS studies are often making their findings publicly available to facilitate collaborative efforts (e.g., the LD Hub; http://ldsc.broadinstitute.org). Similarly, personality researchers should begin publishing item-level raw data and outcomes, as this will facilitate the identification of item-level associations and predicting outcomes across multiple studies. For example, what we did across subsamples could be done across studies. At least, item-level association profiles should be published, so that they could be recycled (e.g., for predicting not-yet-measured outcomes) or meta-analyzed in other studies (e.g., Mõttus et al., under review).

Here, if only implicitly, we have treated outcomes as dependent variables—something to which personality may potentially contribute to. Of course, what we conceptualize as personality may often partly result from variability in the outcomes such as educational qualification or occupational level, or they may spuriously correlate due to shared causal factors such as polygenic influences (Turkheimer, Pettersson, & Horn, 2014; Mõttus, Realo et al., 2017). However, regardless of the direction of the causality between personality and outcomes, representing their associations as accurately as possible is likely to contribute to a better understanding of them. Also, we should note that most associations were longitudinal in this study, with personality being measured about five years before outcomes: in some cases, this may have diminished the probability of outcomes (e.g., 'Sleeping enough') "leaking" into personality ratings.

Conclusion

We report that predictive models based on 50 items, treated as markers of personality nuances, tend to explain more variance in a wide range of outcomes than models based on the Big Five domains. On average, the predictive advantage was an admittedly modest 1.3%. Should anyone care? We think that there are reasons to heed these findings. First, although the difference between more parsimonious domain-level models and more complex item-level models is not large in absolute terms—most effect sizes are small to start with and probably for a good reason—outcomes are multiply determined. Therefore, the relative difference is more substantial: moving from domain- to item-models confers a 30% increase in the amount of variance accounted for in outcomes. Second, this is likely to be a lower-bound estimate of the incremental value of item-models, given that the instrument used in this study was short and limited in item content, and it was deliberately designed to measure the Big Five traits and nothing else. With more comprehensive item pools, the incremental predictive value of item-models will probably be larger. Third, our analyses lend credit to the hypothesis that personality-outcome associations, even if they are modeled using more parsimonious domains, are driven by the specific personality characteristics that individual items are markers of, either directly or by means of being in "linkage" with the directly relevant characteristics. Therefore, even if and when domains do allow for a reasonable prediction of outcomes, they may often not be useful as *explanatory* units for the associations. In order to move from correlations towards explanations, item-models may turn out to be more helpful in the end. To the extent that causal contributions from personality to outcomes are plausible at all, we now have evidence that outcomes may be highly polycausal (polynuanced) and that this is especially plausible for outcomes that are more strongly linked with personality. These findings alone are informative. In conclusion, where sample sizes are large enough, future personality research could routinely build prediction models from items, in addition to domain-based models. Importantly, this comes at *no additional cost* in terms of data collection. Ultimately, more accurate descriptions can help with more realistic explanations.

References

Allik, J., Realo, A., Mõttus, R., Borkenau, P., Kuppens, P., & Hřebíčková, M. (2010). How people see others is different from how people see themselves: A replicable pattern across cultures. *Journal of Personality and Social Psychology*, *99*, 870-882. https://doi.org/10.1037/a0020963

Anglim, J., & Grant, S. (2016). Predicting psychological and subjective well-being from personality: Incremental prediction from 30 facets over the Big Five. *Journal of Happiness Studies*, *17*, 59-80. https://doi.org/10.1007/s10902-014-9583-7

Asendorpf, J. B., Baumert, A., Schmitt, M., Blum, G., van Bork, R., Rhemtulla, M., … & Mõttus, R. (2016). Open peer commentary and author's response. *European Journal of Personality*, *30*, 304–340. https://doi.org/10.1002/per.2060

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., … Wrzus, C. (2017). Integrating personality structure, personality process, and personality

development. *European Journal of Personality*, *31*, 503–528. https://doi.org/10.1002/per.2115

Brown, M., Elliott, J., Hancock, M., Shepherd, P., & Dodgeon, B. (2012). National Child Development Study 2008-2009 Follow-up. *London: CLS, Institute of Education.*

Brown, M., & Hancock, M. (2015). National Child Development Survey 2013 Follow-up: A guide to the datasets. *London: CLS, Institute of Education.*

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., ... & Daly, M. J. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics, 47*, *1236-1244*. https://doi.org/10.1038/ng.3406

Carlo, G., Okun, M. A., Knight, G. P., & de Guzman, M. R. T. (2005). The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation. *Personality and Individual Differences*, *38*, 1293-1305. https://doi.org/10.1016/j.paid.2004.08.012

Cesarini, D., & Visscher, P. M. (2017). Genetics and educational attainment. *Npj Science of Learning*, *2*(1), 4. https://doi.org/10.1038/s41539-017-0005-6

Chabris, C. F., Lee, J. J., Cesarini, D., Benjamin, D. J., & Laibson, D. I. (2015). The fourth law of behavior genetics. *Current Directions in Psychological Science*, *24*, 304-312. https://doi.org/10.1177/0963721415580430

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods*, *21*, 603-620. https://doi.org/10.1037/met0000088

Christiansen, N. D., & Robie, C. (2011). Further consideration of the use of narrow trait scales. *Canadian Journal of Behavioural Science/Revue Canadienne des Sciences du Comportement*, *43*, 183-194. https://doi.org/10.1037/a0023069

Condon, D.M., Roney, E. & Revelle, W., (2017). A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. J*ournal of Open Psychology Data. 5,* 3. http://doi.org/10.5334/jopd.32

Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., … & Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*, 414–431. https://doi.org/10.1002/per.1866

Damian, R. I., Su, R., Shanahan, M., Trautwein, U., & Roberts, B. W. (2015). Can personality traits and intelligence compensate for background disadvantage? Predicting status attainment in adulthood. *Journal of Personality and Social Psychology, 109*, 473–489. https://doi.org/10.1037/pspp0000024

Davies, G., Armstrong, N., Bis, J. C., Bressler, J., Chouraki, V., Giddaluru, S., ... & Van Der Lee, S. J. (2015). Genetic contributions to variation in general cognitive function: A meta-analysis of genome-wide association studies in the CHARGE consortium (N= 53 949). *Molecular Psychiatry*, *20*, 183-192. https://doi.org/10.1038/mp.2014.188

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLOS Genetics, 9,* e1003348. https://doi.org/10.1371/journal.pgen.1003348.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1-22. https://doi.org/10.18637/jss.v033.i01

Goldberg, L. R. (1990). An alternative "description of personality": The Big Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216-1229. https://doi.org/10.1037//0022-3514.59.6.1216

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment, 4,* 26-42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). *The structure of personality traits: Vertical and horizontal aspects.* In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), Studying lives through time: Personality and development (pp. 169–188). Washington, DC: American Psychological Association.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7–28). Tilburg: Tilburg University Press.

Goodwin, R. D., & Friedman, H. S. (2006). Health status and the five-factor personality traits in a nationally representative sample. *Journal of Health Psychology*, *11*, 643-654. https://doi.org/10.1177/1359105306066610

Gough, H. G. (1975). *Manual for the California Psychological Inventory*. Consulting Psychologists Press, Palo Alto, CA.

Graham, E. K., Rutsohn, J. P., Turiano, N. A., Bendayan, R., Batterham, P. J., Gerstorf, D., … & Mroczek, D. K. (2017). Personality predicts mortality risk: An integrative data analysis of 15 international longitudinal studies. *Journal of Research in Personality, 70,* 174-186. https://doi.org/10.1016/j.jrp.2017.07.005

Hathaway, S. R. and McKinley, J. C. (1940). A Multiphasic Personality Schedule (Minnesota) : I. Construction of the Schedule. *The Journal of Psychology, 10,* 249–254. https://doi.org/10.1080/00223980.1940.9917000

Hill, W. D., Davies, G., McIntosh, A. M., Gale, C. R., & Deary, I. J. (2017). A combined analysis of genetically correlated traits identifies 107 loci associated with intelligence. *BioRxiv*, 160291. https://doi.org/10.1101/160291

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*, 95-108. https://doi.org/10.1038/nrg1521

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55-67. https://doi.org/10.1080/00401706.1970.10488634

Jang, K. L., McCrae, R. R., Angleitner, A., Riemann, R., & Livesley, W. J. (1998). Heritability of facet-level traits in a cross-cultural twin sample: Support for a hierarchical model of personality. *Journal of Personality and Social Psychology*, *74*, 1556-1565. https://doi.org/10.1037/0022-3514.74.6.1556

Jokela, M., Batty, G. D., Nyberg, S. T., Virtanen, M., Nabi, H., Singh-Manoux, A., & Kivimäki, M. (2013). Personality and all-cause mortality: Individual-participant meta-analysis of 3,947 deaths in 76,150 adults. *American Journal of Epidemiology, 178,* 667-675. https://doi.org/10.1093/aje/kwt170

Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, antisocial behavior, and aggression: A meta-analytic review. J*ournal of Criminal Justice, 39,* 329–337. https://doi.org/10.1016/j.jcrimjus.2011.03.004

Laidra, K., Pullmann, H., & Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences*, *42*, 441–451. https://doi.org/10.1016/j.paid.2006.08.001

Lee, S. H., DeCandia, T. R., Ripke, S., Yang, J., Schizophrenia Psychiatric Genome-Wide Association Study Consortium (PGC-SCZ), International Schizophrenia Consortium (ISC), ... & Wray, N.R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, *44*, 247-250. https://doi.org/10.1038/ng.1108

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., ... & Croteau-Chonka, D. C. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*, 197-206. https://doi.org/10.1038/nature14177

Major Depressive Disorder Working Group of the PGC, Wray, N. R., & Sullivan, P. F. (2017). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *BioRxiv*, 167577. https://doi.org/10.1101/167577

Malouff, J. M., Thorsteinsson, E. B., Rooke, S. E., & Schutte, N. S. (2007). Alcohol involvement and the Five-Factor Model of personality: A meta-analysis. *Journal of Drug Education*, *37*, 277-294. https://doi.org/10.2190/DE.37.3.d

Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2006). The five-factor model of personality and smoking: A meta-analysis. *Journal of Drug Education*, *36*, 47-58. https://doi.org/10.2190/9EP8-17P8-EKG7-66AD

Malouff, J. M., Thorsteinsson, E. B., Schutte, N. S., Bhullar, N., & Rooke, S. E. (2010). The five-factor model of personality and relationship satisfaction of intimate partners: A meta-analysis. *Journal of Research in Personality*, *44*, 124-127. https://doi.org/10.1016/j.jrp.2009.09.004

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*, 97-112. https://doi.org/10.1177/1088868314541857

McCrae, R. R., & Costa, P. T. (2008). *Empirical and theoretical status of the five-factor model of personality traits.* In B. Boyle, G. Matthews, & D. Saklofske (Eds.), The SAGE handbook of personality theory and assessment: Volume 1 — Personality theories and models (pp. 273–295). London: SAGE.

McCrae, R. R., & Costa, P. T. (2010). *NEO Inventories Professional Manual*. Odessa, FL: Psychological Assessment Resources.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*, 175-215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244. https://doi.org/10.2466/PR0.66.1.195-244

Mõttus, R. (2016). Towards more rigorous personality trait–outcome research. *European Journal of Personality*, *30*, 292–303. https://doi.org/10.1002/per.2041

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 112*, 474-490. https://doi.org/10.1037/pspp0000100

Mõttus, R., Marioni, R., & Deary, I. J. (2017). Markers of psychological differences and social and health inequalities: Possible genetic and phenotypic overlaps. *Journal of Personality*, *85*, 104–117. https://doi.org/10.1111/jopy.12220

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality*, *52*, 47-54. https://doi.org/10.1016/j.jrp.2014.07.005

Mõttus, R., Realo, A., Allik, J., Deary, I. J., Esko, T., & Metspalu, A. (2012). Personality traits and eating habits in a large sample of Estonians. *Health Psychology*, *31*, 806–814. https://doi.org/10.1037/a0027041

Mõttus, R., Realo, A., Allik, J., Esko, T., Metspalu, A., & Johnson, W. (2015). Within-trait heterogeneity in age group differences in personality domains and facets: Implications for the development and coherence of personality traits. *PLoS ONE, 10,* e0119667. https://doi.org/doi:10.1371/journal.pone.0119667

Mõttus, R., Realo, A., Vainik, U., Allik, J., & Esko, T. (in press). Educational attainment and personality are genetically intertwined. *Psychological Science.* https://doi.org/10.1177/0956797617719083

Mõttus, R., Sinick, J., Terracciano, A., Hrebickova, M., Kandler, C., Ando, J., … Jang, K. L. (under review). Personality characteristics below facets (meta-analysis). Retrieved from osf.io/wjmb3. https://doi.org/10.17605/OSF.IO/WJMB3

Neale, M. C., & Maes, H. H. M. (1996). *Methodology for Genetic Studies of Twins and Families.* Dordrecht, The Netherlands: Kluwer Academic Publishers B.V.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., … Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, *533*, 539–542. https://doi.org/10.1038/nature17671

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401-421. https://doi.org/10.1146/annurev.psych.57.102904.190127

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*, 524-539. https://doi.org/10.1037/0022-3514.81.3.524

Plewis, I., Calderwood, L., Hawkes, D., & Nathan, G. (2004). Changes in the NCDS and BCS70 populations and samples over time. *London: CLS, Institute of Education*.

Plomin, R., & Stumm, S. von. (2018). The new genetics of intelligence. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg.2017.104

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*, 322–338. https://doi.org/10.1037/a0014996

R Development Core Team. (2016). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revelle, W. (2017) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA. Retrieved from http://CRAN.R-project.org/package=psych.

Rhodes, R. E., & Smith, N. E. I. (2006). Personality correlates of physical activity: A review and meta-analysis. *British Journal of Sports Medicine*, *40*, 958-965. https://doi.org/10.1136/bjsm.2006.028860

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*, 313-345. https://doi.org/10.1111/j.1745-6916.2007.00047.x.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., Leeuw, C. A. de, … Posthuma, D. (2017). GWAS meta-analysis (N=279,930) identifies new genes and functional links to intelligence. *BioRxiv*, 184853. https://doi.org/10.1101/184853

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113,* 117-143. https://doi.org/10.1037/pspp0000096

Sutin, A. R., Ferrucci, L., Zonderman, A. B., & Terracciano, A. (2011). Personality and obesity across the adult life span. *Journal of Personality and Social Psychology, 101,* 579–592. https://doi.org/10.1037/a0024286

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267-288.

Turkheimer, E., Pettersson, E., & Horn, E. E. (2014). A phenotypic null hypothesis for the genetics of personality. *Annual Review of Psychology*, *65*, 515–540. https://doi.org/10.1146/annurev-psych-113011-143752

University of London. Institute of Education. Centre for Longitudinal Studies. (2012). *National Child Development Study: Sweep 8, 2008-2009*. [data collection]. *3rd Edition.* UK Data Service. SN: 6137, https://doi.org/10.5255/UKDA-SN-6137-2

University of London. Institute of Education. Centre for Longitudinal Studies. (2014). *National Child Development Study: Childhood Data, Sweeps 0-3, 1958-1974*. [data collection]. *3rd Edition.* National Birthday Trust Fund, National Children's Bureau, [original data producer(s)]. UK Data Service. SN: 5565, https://doi.org/10.5255/UKDA-SN-5565-2

University of London. Institute of Education. Centre for Longitudinal Studies. (2015). *National Child Development Study: Sweep 9, 2013*. [data collection]. UK Data Service. SN: 7669, https://doi.org/10.5255/UKDA-SN-7669-1

Vainik, U., Mõttus, R., Allik, J., Esko, T., & Realo, A. (2015). Are trait–outcome associations caused by scales or particular items? Example analysis of personality facets and BMI. *European Journal of Personality*, *29*, 622-634. https://doi.org/10.1002/per.2009

Vecchione, M., Schoen, H., Castro, J. L. G., Cieciuch, J., Pavlopoulos, V., & Caprara, G. V. (2011). Personality correlates of party preference: The Big Five in five big European countries. *Personality and Individual Differences*, *51*, 737-742. https://doi.org/10.1016/j.paid.2011.06.015

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, *4,* 270. https://doi.org/DOI: org/10.3389/fgene.2013.00270

Weiss, A., Gale, C. R., Batty, G. D., & Deary, I. J. (2013). A questionnaire-wide association study of personality and mortality: The Vietnam Experience Study. *Journal of Psychosomatic Research*, *74*, 523-529. https://doi.org/10.1016/j.jpsychores.2013.02.010

Wittmann, W. W. (1988). *Multivariate reliability theory. Principles of symmetry and successful validation strategies.* In J. R. Nesselroade & R. B. Cattell (Eds.), Handbook of multivariate experimental psychology (pp. 506–560). New York: Plenum Press.

Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, *122*, 84–11. https://doi.org/10.1037/a0038423

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122. https://doi.org/10.1177/1745691617693393

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x